A Journey Through the Reliability of a Decision-Making Model for Testing Students With Disabilities

Gerald Tindal

Leanne Ketterlin-Geller

University of Oregon

Address all correspondence to: Gerald Tindal, Castle-McIntosh-Knight Professor Education, 270 Educational Leadership, 5267 University of Oregon, Eugene, OR, 97403-5267

A Journey Through the Reliability of a Decision-Making Model for Testing Students With Disabilities

A need exists for accommodations in large-scale testing, though it is somewhat uncertain how and why they appear to be effective with which students. The empirical support for them is inconsistent within and across subject areas and even though they sometimes to be effective, they also appear to be either inert (not work for anyone) or overly effective (work for everyone). It is not yet possible, therefore, to simply move research to practice in adopting wide scale adoption of specific accommodations that have passed the test of replicable empirical support. Yet, large-scale testing requires their application. To bridge this gap, the research on accommodations has begun to focus on how teachers make the decision to recommend specific changes in testing.

Teacher Decision-Making on Accommodations

The need exists to understand practice in teacher making for no other reason than that we know so little about it. Indeed, the early findings reported by Hollenbeck, Tindal, and Almond (1998) indicates that only 55% of the general and special education teachers are correct in determining whether or not an accommodation is allowable. Even more problematic is the finding that special and general education teachers are not different from each other according to these authors.

Though not directly focused on accommodations, Crawford and Tindal (2002) studied teachers' perceptions of the participation of students with disabilities in large-scale testing by organizing their comments into three categories: (a) teacher knowledge, (b) teacher attitude (in which the comments were not as abundant but were quite emotional), and (c) teacher decision-making (which contained the majority of comments). In this last category, they reported similar findings to those reported by Jayanthi, Epstein, Polloway, and Bursuck (1996): Many decisions about participation in large-scale tests are made by individual teachers not by Individualized Educational Program (IEP) teams. Individual student characteristics (and basic skills) were the primary reference in making these decisions. As they conclude: "Teachers should be trained to use student performance data to validate these [inclusion] decisions…special service providers should develop a firm understanding of test accommodations available to students with disabilities" (p. 114). As two teachers so eloquently stated the problem in the Crawford and Tindal (2002) study: "I think we need to be trained and more information should be disbursed for us" and "We have some accommodations and some modifications but it looks like it's not clear how far we can push the envelope" (p. 107).

Using a similar focus group methodology, Ysseldyke, Thurlow, Bielinski, House, Moody, & Haigh, (2001) investigated the alignment of test accommodations with those used in instruction (specifically IEPs): "If a student had an IEP goal, it was very likely that the student received an accommodation for instruction in that area" (p. 216). Indeed, 82% of the students in their sample received some form of accommodation though no differences were found by disability prevalence or type. Importantly, 84% had instructional accommodations that matched their testing accommodations. Though they distinguished between accommodations and modifications, it appeared that this distinction was based solely on the orientation to the standards, as reading the reading test was viewed as an accommodation.

Given that teachers may or may not even be knowledgeable about allowable accommodations and with the press to ensure that accommodations in their classrooms are consistent with those used in the testing situation, it is important to support teachers decision-making practices at the same time as basic research on accommodations is proceeding. This kind of support must come from supplemental information that is collected in addition to the purely descriptive information on state test results for two reasons. First, such outcome data usually represent post hoc results and teachers need information to make the initial decision. Second, descriptive information on state test results from accommodated and non-accommodated conditions is confounded by student characteristics (non-accommodated students are likely to be a different population of students than those who have been recommended to receive an accommodation).

Four systems have emerged for understanding teacher decision-making on accommodations, differing primarily on the source of data that they use. For Fuchs and colleagues, the focus has been on using curriculum-based measurement (CBM) as companion data for making decisions about accommodations and their effects. Basically, teachers administer a basic skills measure in reading or mathematics to make a prediction about the need for an accommodation; in their research designs, this prediction is compared to those made by teachers using informal information. For Elliott and colleagues, the source of information is a checklist on accommodations that help structure teachers' rationale for recommending accommodations. DeStephano and colleagues focus on students' Individualized Educational Programs (IEPs) to ascertain the need for accommodations (and consistency with instructional use). Finally, Tindal, Ketterlin-Geller and colleagues use CBM as part of a diagnostic prediction that can be confirmed by documenting the effects of accommodations. Following are some specific findings from these four systems for recommending accommodations.

Fuchs and Colleagues

"One major obstacle to valid participation is the lack of standard methods for determining which testing accommodations preserve the meaningfulness of scores (Fuchs & Fuchs, 2001, p. 174). Because the research base is thin, the population of students with disabilities is heterogeneous, and teachers have difficulty making recommendations when using informal judgments, they propos making data based decisions. Their system – Dynamic Assessment of Test Accommodations – is designed to assist teachers in making recommendations for test accommodations that include extended time, reading problems aloud (in math), use of calculators, an adult writing non-mathematical responses, and large print. Accommodations are recommended by comparing a student's boost to that which can be expected (based on normative information from a population of students with learning disabilities).

In comparing accommodations recommended in this system with those recommended by teachers (or assigned at random), they reported significant differences: "Students to whom DATA had awarded accommodations earned larger boosts as a function of having those accommodations, compared to the subset to whom DATA had denied accommodations. The effect size was 0.34 standard deviations" (p. 179). Teachers both awarded and denied accommodations in a manner that reflected false positives and false negatives.

Elliott and Colleagues

Schulte, Elliott, and Kratochwill (2000) used case vignettes to study the selection of assessment accommodations using a research design that allowed them to study the nature of the disability and the type of the assessment task. Using the Assessment Accommodations Checklist (a checklist with 74 accommodations divided among eight categories and rated on use, potential helpfulness, and fairness, they described their purpose as examining "educators' perceptions of the MC as a tool for generating accommodation ideas and then documenting and evaluating assessment accommodations used with students" (p. 47). They reported five findings:

1. No differences existed in the selection of accommodations for students with significant disabilities versus learning disabilities.

2. Accommodations were judged as equally helpful for both these student populations.

3. More accommodations were selected for performance assessments than selection-response assessments (e.g., multiple-choice test).

4. Some recommended accommodations were rated as more helpful and fair for performance assessments.

5. The ACC was deemed to be a relevant and useful tool.

DeStephano-Shriner and Colleagues

DeStephano, Shriner, and Lloyd (2001) developed a model for training teachers on decision-making for participation in large-scale assessments that was based on present levels of performance in their IEPs. Working from the perspective that assessment accommodations should be parallel with those used in instruction (using the IEP as a proxy for instruction) and assuming that accommodations should be implemented to "mediate the effects of 'access' deficits but not invalidate the assessment of 'target' skills" (p. 9), they created six scenarios for participation and trained teachers how to make decisions about accommodations. In their training, they included information about IDEA requirements, IEP modifications, familiarity with content standards, and a flow chart illustrating how IEPs could be used for accommodation and participation decisions. Finally, they considered both the participation of the student in the general curriculum, the use of accommodations, and the roles of both general and special education teachers. They reported significant changes in the participation rates and accommodation patterns as a result of their training and in relation to accessing the general curriculum with appropriate accommodations. "After training, teachers' decisions about assessment participation and accommodation did show a stronger link to students' access to the general curriculum and needed instructional accommodations than decisions prior to training. Accommodations for target skills are markedly reduced" (p. 18).

Tindal, Ketterlin-Geller and Colleagues

This group of researchers has approached the process for recommending accommodations with a computer-based accommodation station (AS) in which a series of basic skills assessments are administered and perceptions are documented with a report generated for IEP teams to use in making a recommendation. The AS has the following measures embedded in the computer programming.

1. A silent reading measure is used in which students are directed to move to the next screen where a passage is presented for them to read and, when done, 'click next'. When the screen is first presented, a clock begins timing the student and it is stopped when the student 'clicks next'. This measure is based on the work of Miller (1990) and Yule (1987) and is described by Ketterlin-Geller, Alonzo, Carrizales, & Tindal (2005); it allows the rate of silent reading to be calculated.

2. A series of sentences are presented for a fixed period of time, each of them replaced by a blank screen, which is then followed by a literal comprehension question addressing information from the sentence. Students with reading problems often click immediately to move off the passage (presumably because they can not read the full passage); this task allows calculation of correct responses on very easy-to-read text and easy-to-answer comprehension questions.

3. A maze test measures student comprehension and is based on the technical work summarized by Shin, Deno, and Espin (2000). In this task, a passage is presented with the first sentence left intact and thereafter every nth (frequently every 7th) word is deleted and the student

is directed to select one of four options to correctly complete the sentence. This measure includes 16 fill-in-the-blanks to document the student's understanding of syntax, grammar, and semantics.

4. Various mathematics problems are presented in both accommodated and standard fashion to determine whether or not the student's performance is differentially affected. These accommodations include simplification of language, reading the mathematics problem aloud, and presenting the problem in Spanish – see Ketterlin-Geller, Yovanoff, and Tindal (in press).

5. A series of statements are presented that address student skills, interests, and benefit from various changes to the testing situation. Teachers and students respond on a scale of agreement, representativeness, or likelihood. These items reflect the field-testing work conducted by Alonzo, Ketterlin-Geller, and Tindal (2004).

Summary of Teacher-decision Making on Accommodations

The four models for making accommodations recommendations vary primarily in the data sources that are used and may vary in their technical adequacy. At this point, the CBMs from the Fuchs look very promising, the accommodations checklists from Elliott appear very popular, the focus on IEPs by Destaphano highly relevant, and the Accommodation Station potentially useful for IEP teams. Yet, further research is needed on all of them.

As Bolt and Thurlow (2004) recommend, the following practices should be followed:

1. Make the skills explicit prior to making accommodations decisions.
2. Use the least intrusive accommodations.
3. Align assessment with instruction.
4. Train test administrators in implementation of the accommodation.
5. Anticipate difficulties and be prepared to address challenges.
6. Monitor accommodations outcomes for individual students.

It is quite likely that the experimental research on accommodations needs to move to a field-based platform that both allows teachers to make decisions and systematically investigates the effects using randomized designs. In this process, more careful analysis of the achievement construct is needed at the item level and more rich descriptions are needed of the populations being tested.

Methods

The overall goal of the AS pilot study currently underway is to investigate the reliability and utility of the Accommodation Station (AS), an online decision-making model that helps IEP teams determine which testing accommodations are appropriate for individual students with disabilities. The Accommodation Station pilot study took place in South Carolina, Maryland, and Pennsylvania during November, December, and January. Testing in Oregon began in March. Approximately 60 students with learning disabilities and current IEPs in each of grades 3, 5, and 8 will participate in each state, for a total N of 180 students per grade in SC, MD, and PA; in Oregon, 100 students per grade were tested.

The Accommodation Station pilot study included a student, teacher, and parent component. Students participated in two ninety-minute online administrations of the Web-based Accommodation Station. During each administration of the AS, students completed a short section of reading and math items and responded to survey questions about their learning and testing preferences.

In addition, two teachers per student responded to a set of online survey questions about their students' skills and abilities, as well as the instructional strategies and accommodations they employed with individual students. One general education and one special education teacher responded to the survey questions about each individual student. The parents/guardians of participating students also answered a similar set of questions about their child. The paper-based parent survey was enclosed with the parent notification letter sent home with participating students. Students, teachers, and parents took the AS/filled out the surveys twice within a two-week window between administrations. The test-retest design of this pilot study allowed us to determine if the AS was a reliable tool.

The following materials for the Accommodation Station pilot studies were drafted and sent to partner states: sample district superintendent letter, sample parent notification letter, AS parent survey, a list of teacher roles and responsibilities, and talking points on the AS for recruiting schools. Partner state contracts also were drafted and sent to partner states for review. In addition, these materials also were drafted and sent to participating pilot study schools: a technical checklist and manual for teachers (to ensure computer labs and computers can support the AS), a template for entering student names and background variables into the AS, and a list of student variables for schools to refer to in filling out the template. A principal letter was drafted and sent to participating schools.

Two test runs of the AS were conducted in South Carolina prior to the larger-scale pilot studies. These run-out studies provided insights into the revised system and some adjustments were made before the larger pilot studies began in South Carolina and the partner states.

Some Initial Findings from the Reading Tasks (SC Only)

Maze

1. Most students took the first maze (and didn't get the boot because of not responding or random responding):

    a. Session ONE – Time 1 = 90% and Time 2 = 8%.

    b. Session TWO – Time 1 = 55% and Time 2 = 12%.

2. The performance of students who got the boot is lower than those who took it procedurally correct the first time (6 vs. 4 in Session ONE and 7 versus 3 in Session TWO).

3. The maze is difficult (averaging 6 of 16 items correct across the grades; it is lowest in grade 5 (4 of 16 items correct) and grade 3 (5 of 16 items correct); in grade 8, performance averages 8 of 16 items correct. Performance is consistent from Session ONE to Session TWO.

4. Students switch their responses on the maze quite frequently from Session ONE to Session TWO about half the time.

Repeated Reading

1. When presented a sentence to read (*Grandma's house* had 10 words, *Sue's invitation* had 11 words, *Birds* had 11 word, and *Hiding at Ann's* had 15 words), and then a blank screen followed by a question to answer about the sentence with four options. On this task, students performed quite poorly.

2. In Session ONE:
    a. In grade 3 (n=23), the average difficulty of four questions ranged from .26 to .52.
    b. In grade 5 (n=31), the range of difficulty of the four questions was .26 to .42.
    c. In grade 8 (n=65), this range of difficulty for the four questions was .46 to .78.

3. In Session TWO:
    a. In grade 3 (n=12), the average difficulty of four questions ranged from .50 to .83.
    b. In grade 5 (n=20), the range of difficulty of the four questions was .30 to .50.
    c. In grade 8 (n=42), this range of difficulty for the four questions was .69 to .86.

4. The correlation between the repeated reading correct performances across sessions was very low at .21.

Silent Reading Fluency

1. In Session ONE:
    a. In grade 3 (n=17-22), the average fluency ranged from 68 WPM to 83 WPM.
    b. In grade 5 (n=27-31), the average fluency ranged from 73 WPM to 77 WPM.
    c. In grade 8 (n=59-61), average fluency ranged from 93 WPM to 112 WPM.
2. In Session TWO:
    a. In grade 3 (n=8-10), the average fluency ranged from 84 WPM to 89 WPM.
    b. In grade 5 (n=19-20), the average fluency ranged from 81 WPM to 93 WPM.
    c. In grade 8 (n=30-w32), average fluency ranged from 144 WPM to 148 WPM.
3. The correlation between these passages (WPM) was moderate: .64 between passages in Session ONE and .76 between passages in Session TWO. The correlations across sessions were low: .20, .22, .37, and .47.

Surveys

In surveying students about specific judgments, they were fairly consistent overall, but on specific items were either very consistent or somewhat inconsistent. All statements were consistently judged (with 67% exact agreement or better) except the following that focused on:

1. *How often do you get to do the following things on math tests*:

- Have someone read the problems and directions to you. (33% exact match)
- How often do you get to do the following things on math tests? (31% exact match).
- Have the words in the problems and directions made easier to understand. (31% exact match).
- Have the questions written in the language you speak, like Spanish, or Chinese. (43% exact match).

- Get more time to finish. (51% exact match).
- Take the test as several short tests instead of all at once. (37% exact match).
- Take the test alone somewhere away from the rest of the class. (26% exact match).
- Answer by typing, pointing, or having someone write what you tell them. (29% exact match).
- Have the letters and pictures in the test bigger. (23% exact match).
- Write answers to the questions in your own way. (43% exact match).
- Choose an answer to the questions from a group of choices. (34% exact match).
- Use a calculator. (43% exact match).

2. How easy is it for you to…

- Work on your own for 45-60 minutes? (37% exact match).
- Work as part of the whole class group? (34% exact match).
- Read and understand directions? (46% exact match).
- Take short quizzes? (49% exact match).
- Take long tests? (51% exact match).
- Take tests on computers? (43% exact match).

3. How do you feel about…

- Taking state tests? (49% exact match).
- Taking classroom tests? (31% exact match).
- Working with computers? (57% exact match).
- Taking tests on computers? (46% exact match).

The computational math was somewhat consistent from session ONE to session TWO:

- 46% exact agreement for problem 1 (with only 11 students).
- 100% exact agreement for problem 2 (with only 7 students).
- 56% exact agreement for problem 3 (with only 9 students).
- 56% exact agreement for problem 4 (with only 9 students).

Discussion

The preliminary findings from this study indicate that decision-making for accommodations is very difficult to reliably complete (using student results) and reveals mixed results: While student comprehension (maze performance) is fairly unreliable, student rapid and silent reading may be more reliable. Nevertheless, student perceptions are quite unreliable. At this time, the reliability of teacher perceptions is being analyzed.

The source of such unreliability in the reading access skills is uncertain. In part, performance of students could be a function of the computer-based nature of the tasks and students' familiarity with them. For example, the rapid reading task is somewhat unusual (and apparently is somewhat difficult) with students not being presented with this kind of task in everyday, classroom routines. It also, however, may serve as an important access skill that is related to short-term memory. Nevertheless, silent reading, appears quite stable within sessions but may not be stable across sessions (though the sample size was very low). It is completed everyday in the classroom and may be a better indicator of reading access skill than repeated reading brief questions.

Students' unreliability about perceptions, however, is more difficult to explain, particularly their experience with having previously received various accommodations in the classroom. Perhaps their lack of consistency in noting their accommodation is a function of the ‘noticeability’ of the accommodations: Teachers use them in such a manner that it becomes part of the fabric of instruction and students don’t even notice it when asked to reflect on it. It also may be due to the manner in which we labeled the accommodations on the survey: Though students receive a particular accommodation, their teacher never labels it as such. Finally, their lack of consistency may indeed reflect the lack of consistency in receiving it and their responses function from their most recent experience.

Whatever the reason for the marginal reliability (stability) of various skills and perceptions that are relevant for making recommendations for accommodations, much more clear and explicit training is needed. This training may focus on any of the four systems that were reviewed in the introduction. Using Fuchs and Fuchs (2001) system, teachers would be trained on the administration of curriculum-based measures and then its use in making decisions. DeStephano, Shriner, and Lloyd (2001) already have a clear training system that appears to be effective in linking classroom use with use in large-scale testing; it does not, however, help in making the initial recommendation for use in the classroom and does not relate to actual student performance. Elliott’s system (Schulte, Elliott, & Kratochwill 2001) for teachers to follow a checklist is standard practice but, like the IEP analysis model, fails to relate to students’ actual access skills; furthermore, the reliability (stability) of the checklist needs to be verified much the same as noted in this study. Finally, the Accommodation Station itself (Ketterlin-Geller, Tovanoff, & Tindal, in press) may need further study in the manner in which it is packaged and used. More practice items assessing reading skills may be needed in which the student receives feedback; teachers and students may need better training in how to reflect on the various dimensions of accommodations: the importance of various access skills, their use in the classroom, and their benefit in helping students perform on large-scale tests.

References

Alonzo, J., Ketterlin-Geller, L., & Tindal, G. (2004). *Instrument development: Examining the appropriateness of student and teacher surveys for determining the need for testing accommodations.* (No. Technical Report 31). Eugene, OR: Behavioral Research and Teaching: University of Oregon.

Crawford, L., Almond, P., Tindal, G., & Hollenbeck, K. (2002). Teacher perspectives on inclusion of students with disabilities in high-stakes assessments. *Special Services in the Schools, 18*(1/2), 95-118.

DeStefano, L., Shriner, J. G., & Lloyd, C. A. (2001). Teacher decision making in participation of students with disabilities in large-scale assessment. *Exceptional Children, 66*(1), 7-22.

Fuchs, L. S., Fuchs, D. (2001). Helping teachers formulate sound test accommodation decisions for students with learning disabilities. *Learning Disabilities Research & Practice, 16*(3), 174-181.

Hollenbeck, K., Tindal, G., & Almond, P. (1998). Teachers' knowledge of accommodations as a validity issue in high-stakes testing. *The Journal of Special Education, 32*(3), 175-183.

Jayanthi, M. E., M. H., Polloway, E. A., & Brusuck, W. D. (1996). A national survey of general education teachers' perceptions of testing adaptations. *The Journal of Special Education, 30*(1), 99-115.

Ketterlin-Geller, L., Alonzo, J., Carrizales, D., & Tindal, G. (2005). *Reading between the lines: Testing methods of capturing silent reading fluency.* Eugene, OR: University of Oregon.

Ketterlin-Geller, L., Yovanoff, P., & Tindal, G. (in press). Developing a new paradigm for conducting research on accommodations in mathematics testing. *Exceptional Children*.

Miller, S. D., & Smith, D. E. P. (1990). Relations among oral reading, silent reading and listening comprehension of students at differing competency levels. *Reading Research and Instruction, 29*, 73-84.

Schulte, A. A., Elliott, S. N., & Kratochwill, T. R. (2001). *Effects of testing accommodations on standardized mathematics test scores: An experimental analysis of the performance of students with and without disabilities.* Madison, WI: Wisconsin Center for Education Research.

Shin, J., Deno, S. L., & Espin, C. (2000). Technical adequacy of the maze task for curriculum-based measurement of reading growth. *The Journal of Special Education, 34*(3), 164-172.

Ysseldyke, J., Thurlow, M., Bielinski, J., House, A., Moody, M., & Haigh, J. (2001). The relationship between instructional and assessment accommodations in an inclusive state accountability system. *Journal of Learning Disabilities, 34*, 212-220.

Yule, V. (1987). Assessing children's silent reading. *Educational Researcher, 29*, 192-196.